A. Derivation of Eq. (7) and Eq. (15)

A.1. Eq. (7): Null-space equivalence

Let $\mathbf{H}_b \in \mathbb{R}^{d\times N_b}$ be the benign activation matrix. Eq. (7) claims that the left null space of $\mathbf{H}_b$ coincides with that of its Gram matrix:

$$\text{Null}(\mathbf{H}_b) = \text{Null}(\mathbf{H}_b\mathbf{H}_b^\top). \tag{17}$$

The null spaces are defined by:

$$\text{Null}(\mathbf{H}_b) = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^\top\mathbf{H}_b = \mathbf{0}^\top\}, \tag{18}$$

$$\text{Null}(\mathbf{H}_b\mathbf{H}_b^\top) = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^\top\mathbf{H}_b\mathbf{H}_b^\top = \mathbf{0}^\top\}. \tag{19}$$

To show the inclusion $\text{Null}(\mathbf{H}_b) \subseteq \text{Null}(\mathbf{H}_b\mathbf{H}_b^\top)$, take any $\mathbf{x}$ such that $\mathbf{x}^\top\mathbf{H}_b = \mathbf{0}^\top$. Using this condition:

$$\mathbf{x}^\top\mathbf{H}_b\mathbf{H}_b^\top = (\mathbf{x}^\top\mathbf{H}_b)\mathbf{H}_b^\top = \mathbf{0}^\top, \tag{20}$$

which confirms $\mathbf{x} \in \text{Null}(\mathbf{H}_b\mathbf{H}_b^\top)$. Conversely, assume $\mathbf{x}^\top\mathbf{H}_b\mathbf{H}_b^\top = \mathbf{0}^\top$. Since $\mathbf{H}_b\mathbf{H}_b^\top$ is symmetric and positive semidefinite, its quadratic form satisfies:

$$\mathbf{x}^\top\mathbf{H}_b\mathbf{H}_b^\top\mathbf{x} = \|\mathbf{H}_b^\top\mathbf{x}\|_2^2 \geq 0. \tag{21}$$

Multiplying the nullity condition by $\mathbf{x}$ yields:

$$0 = \mathbf{x}^\top\mathbf{H}_b\mathbf{H}_b^\top\mathbf{x} = \|\mathbf{H}_b^\top\mathbf{x}\|_2^2, \tag{22}$$

which forces $\mathbf{H}_b^\top\mathbf{x} = \mathbf{0}$ and thus $\mathbf{x}^\top\mathbf{H}_b = \mathbf{0}^\top$. Hence $\mathbf{x} \in \text{Null}(\mathbf{H}_b)$. Since each null space is contained in the other, the equivalence in Eq. (17) follows:

$$\text{Null}(\mathbf{H}_b) = \text{Null}(\mathbf{H}_b\mathbf{H}_b^\top). \tag{23}$$

A.2. Closed-form Solution of Eq. (15)

We derive the closed-form solution of the regularized least-squares problem in Eq. (14). Recall that the objective is:

$$\begin{aligned}\tilde{\mathbf{\Delta}}^\star = \arg\min_{\tilde{\mathbf{\Delta}}} \Big(&\|\tilde{\mathbf{\Delta}}\mathbf{P}\mathbf{H}_m - \mathbf{R}\|_F^2 \\ &+ \alpha\|\tilde{\mathbf{\Delta}}\mathbf{P}\|_F^2 + \beta\|\tilde{\mathbf{\Delta}}\mathbf{P}\mathbf{H}_m - \mathbf{V}\|_F^2\Big), \\ &\alpha > 0,\ \beta \geq 0.\end{aligned}$$

For notational simplicity, we introduce:

$$\begin{aligned}\mathbf{X} &:= \mathbf{P}\mathbf{H}_m \in \mathbb{R}^{d\times N_m}, \\ \mathbf{Z} &:= \mathbf{P} \in \mathbb{R}^{d\times d}, \\ \mathbf{Y} &:= \mathbf{R} + \beta\mathbf{V} \in \mathbb{R}^{d\times N_m}, \\ \mathbf{W} &:= \tilde{\mathbf{\Delta}} \in \mathbb{R}^{d\times d}.\end{aligned} \tag{24}$$

Grouping the two alignment terms in (A.2), the objective can be written in the compact form:

$$J(\mathbf{W}) = \|\mathbf{W}\mathbf{X} - \mathbf{Y}\|_F^2 + (\alpha+\beta)\,\|\mathbf{W}\mathbf{Z}\|_F^2, \tag{25}$$

which corresponds to a matrix-valued ridge regression with effective target $\mathbf{Y}$ and regularization coefficient $\alpha + \beta$.

Using $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}\mathbf{A}^\top)$, we expand (25) in trace form:

$$\begin{aligned}J(\mathbf{W}) = &\,\text{tr}\big[(\mathbf{W}\mathbf{X} - \mathbf{Y})(\mathbf{W}\mathbf{X} - \mathbf{Y})^\top\big] \\ &+ (\alpha+\beta)\,\text{tr}\big[(\mathbf{W}\mathbf{Z})(\mathbf{W}\mathbf{Z})^\top\big].\end{aligned} \tag{26}$$

Taking the derivative with respect to $\mathbf{W}$ and using the standard rule:

$$\nabla_{\mathbf{W}}\,\text{tr}(\mathbf{W}\mathbf{A}\mathbf{W}^\top\mathbf{B}) = 2\,\mathbf{B}\mathbf{W}\mathbf{A}, \tag{27}$$

we obtain the gradient:

$$\nabla_{\mathbf{W}}J = 2(\mathbf{W}\mathbf{X} - \mathbf{Y})\mathbf{X}^\top + 2(\alpha+\beta)\,\mathbf{W}\mathbf{Z}\mathbf{Z}^\top. \tag{28}$$

Setting $\nabla_{\mathbf{W}}J = \mathbf{0}$ yields the stationarity condition:

$$(\mathbf{W}\mathbf{X} - \mathbf{Y})\mathbf{X}^\top + (\alpha+\beta)\,\mathbf{W}\mathbf{Z}\mathbf{Z}^\top = \mathbf{0}, \tag{29}$$

which can be rearranged into the linear matrix equation:

$$\mathbf{W}\big(\mathbf{X}\mathbf{X}^\top + (\alpha+\beta)\,\mathbf{Z}\mathbf{Z}^\top\big) = \mathbf{Y}\mathbf{X}^\top. \tag{30}$$

The matrix in parentheses may be rank-deficient because $\mathbf{Z} = \mathbf{P}$ is a projection. We therefore adopt the Moore–Penrose pseudoinverse and obtain the minimum-norm solution:

$$\mathbf{W}^\star = \mathbf{Y}\mathbf{X}^\top\big(\mathbf{X}\mathbf{X}^\top + (\alpha+\beta)\,\mathbf{Z}\mathbf{Z}^\top\big)^+. \tag{31}$$

Substituting back $\mathbf{X} = \mathbf{P}\mathbf{H}_m$, $\mathbf{Z} = \mathbf{P}$, $\mathbf{Y} = \mathbf{R} + \beta\mathbf{V}$, and $\mathbf{W} = \tilde{\mathbf{\Delta}}$ into (31) gives:

$$\tilde{\mathbf{\Delta}}^\star = (\mathbf{R}+\beta\mathbf{V})\mathbf{H}_m^\top\mathbf{P}^\top\Big(\mathbf{P}\mathbf{H}_m\mathbf{H}_m^\top\mathbf{P}^\top + (\alpha+\beta)\,\mathbf{P}\mathbf{P}^\top\Big)^+, \tag{32}$$

which coincides with Eq. (15) in the main paper.

B. Implementation Details

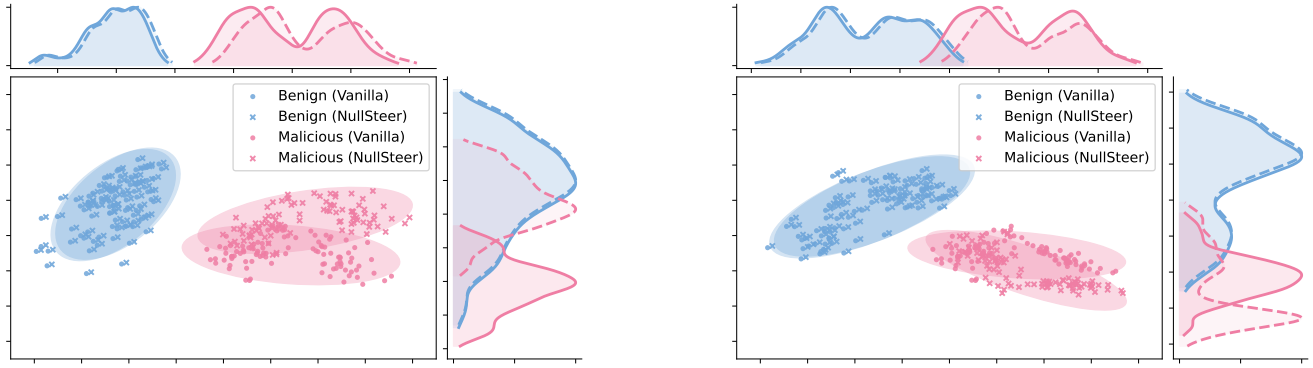
We evaluate our method on three widely adopted open-source VLMs: Qwen2-VL-7B [1], MiniGPT-4-13B [66], and LLaVA-v1.5-13B [28]. Unless otherwise specified, the number of malicious samples used for optimization is set to 96. The steering layer is fixed at l=20 for 13B models and l=14 for the 7B model. For inference, LLaVA-v1.5 and Qwen2-VL use a temperature of 0.2 with top-p=0.9, while MiniGPT-4 follows its official configuration with temperature 1.0 and top-p=0.9.

C. Additional Experimental Results

OOD robustness. Table E evaluates the cross-attack generalization of steering vectors extracted from Jailbreak adversarial samples with $\epsilon = \frac{16}{255}$. NullSteer achieves consistently lower ASR across structured-, perturbation-, and text-only attacks on all three models. On MiniGPT-4, it attains

Table E. Transferability performance under OOD conditions. Steering vectors are obtained from Jailbreak adversarial samples generated with $\epsilon = \frac{16}{255}$, using the same α value determined on the Jailbreak validation split. Their generalization is assessed across unseen attack types, including structure-based attacks from MM-SafetyBench [31], perturbation-based variants of PGD, and text-only attacks. The ASR is computed using the HarmBench [36] classifier.

Methods		Structured-based Attack			Perturbation-based Attack						Text-only Attack
		SD	SD_TYPO	TYPO	PGD [34] $\epsilon = \frac{16}{255}$	PGD [34] $\epsilon = \frac{32}{255}$	Auto-PGD [7] $\epsilon = \frac{16}{255}$	Auto-PGD [7] $\epsilon = \frac{32}{255}$	MI-FGSM [11] $\epsilon = \frac{16}{255}$	MI-FGSM [11] $\epsilon = \frac{32}{255}$	GCG [72]
MiniGPT-4	w/o defense	13.75	43.25	43.75	70.91	78.18	74.55	76.36	78.18	79.09	58.18
	ASTRA	3.75	8.75	11.25	5.45	12.73	5.45	10.91	16.37	13.64	9.09
	NullSteer	2.35	6.28	11.57	3.28	9.43	4.65	9.42	14.21	12.71	8.24
Qwen2-VL	w/o defense	20.00	61.25	38.75	74.55	80.00	76.37	77.57	80.00	78.18	81.82
	ASTRA	11.25	40.00	33.75	21.82	14.55	15.76	15.76	18.18	18.18	30.91
	NullSteer	9.86	38.50	32.53	19.95	13.86	14.28	14.35	16.83	17.28	28.58
LLaVA-v1.5	w/o defense	18.75	55.00	22.50	69.09	74.55	80.60	90.30	87.28	89.09	92.73
	ASTRA	8.75	25.00	6.25	1.82	1.82	1.21	0.61	0.00	0.00	14.55
	NullSteer	7.65	24.00	5.80	1.74	1.70	1.00	0.61	0.00	0.00	13.50



(a) Activation distributions on Qwen2-VL

(b) Activation distributions on LLaVA-v1.5.

Figure I. Comparison of benign and malicious activation distributions across two VLMs.

2.35% on the SD set and 9.09% under GCG, outperforming ASTRA (3.75% and 8.24%) and the undefended model (13.75% and 58.18%). Similar patterns are observed on Qwen2-VL and LLaVA-v1.5, where NullSteer notably reduces attack success under both structured and perturbation-based attacks. These results suggest that the learned steering directions capture transferable semantics of harmful behaviors rather than attack-specific perturbations.

Activation distribution analysis. Figure I presents the activation distributions of benign and malicious inputs on Qwen2-VL and LLaVA-1.5 before and after applying `NullSteer`. Across both models, the benign activations under `NullSteer` closely overlap with their vanilla counterparts, forming compact clusters with nearly unchanged geometry. This demonstrates that the null-space constraint successfully restricts the modification to directions orthogonal to benign semantics, thereby maintaining utility. In contrast, the malicious activations exhibit a pronounced shift away from the benign region after steering, with their density moving consistently toward safer activation directions. The clear separation between the two distributions indicates that `NullSteer` effectively isolates unsafe features while preserving the structure of benign representations, leading to stronger safety without sacrificing model behavior on harmless inputs.